Sci Eng Ethics (2012) 18:633–650
DOI 10.1007/s11948-011-9278-y

And Lead Us (Not) into Persuasion…? Persuasive Technology and the Ethics of Communication

Andreas Spahn

Received: 9 February 2011 / Accepted: 21 April 2011 / Published online: 5 May 2011

Abstract The paper develops ethical guidelines for the development and usage of persuasive technologies (PT) that can be derived from applying discourse ethics to this type of technologies. The application of discourse ethics is of particular interest for PT, since 'persuasion' refers to an act of communication that might be interpreted as holding the middle between 'manipulation' and 'convincing'. One can distinguish two elements of discourse ethics that prove fruitful when applied to PT: the analysis of the inherent normativity of acts of communication ('speech acts') and the Habermasian distinction between 'communicative' and 'strategic rationality' and their broader societal interpretation. This essay investigates what consequences can be drawn if one applies these two elements of discourse ethics to PT.

Keywords Ethics · Persuasive technology · Discourse ethics · Habermas · Speech-act theory · Critical theory · Persuasion

Introduction

In our daily life we are confronted more and more with technologies that try to tell us how we should behave. Our cars annoy us with blinking lights and acoustic signals, when we do not put our safety belts on; our navigation systems complain, when we drive too fast; our computers warn us to regularly update our security software and smart meters try to motivate us to use less energy in our households. These are just a few examples of so-called 'persuasive technologies' (PT). PT are intentionally designed to change the user's attitude, behavior or beliefs (Fogg 2003;

A. Spahn (✉)
Ethics of Technology, School of Innovation Sciences, Eindhoven University of Technology,
Room: IPO 1.02, Postbus 513, 5600 MB Eindhoven, The Netherlands
e-mail: a.spahn@tue.nl

IJsselsteijn 2006) often by giving the user feedback of his actions (or omissions) and by trying to 'suggest' to him a desired pattern of behavior.

A good example of a persuasive technology is the re-design of the dashboard used in the Honda Insight hybrid car. The engineers have completely re-designed the dashboard to help the user to drive more environmentally friendly. Classical car dashboards give mainly feedback on driving speed and fuel tank status. The new 'eco assist' dashboard still has these features, but many persuasive elements have been added. The dashboard has a little display field, on which 'leaves icons' virtually grow, if the user is driving in an environmentally friendly manner. If he, however, accelerates and breaks a lot, and thus wastes a lot of fuel, these leaves will disappear again. Also the speed display changes its background color to signal sustainable or less sustainable driving behavior.

Another example is the Wattson Energy Meter. This is a device that measures the energy consumption in a household and gives feedback via ambient light and via a display. The light changes from blue to red, if the energy consumption increases. The display shows the amount of power in kilowatt hour and the price you have to pay for this energy consumption. These features are meant to raise awareness on energy consumption and train the user to consume less energy.

Advocates of these types of technology point out the noble task of PT. Technology is no longer a neutral tool (if it ever was one), but helps to achieve moral goals like health, safety, sustainability and the like. Critiques argue, however, that this way of 'moralizing' technology raises many ethical concerns. It might even be argued that PT can be regarded as the implementation of a technological paternalism, which conflicts with the ideal of a free and autonomous choice of the individual. Since PT do not *convince* the user to change his behavior or attitudes, but *persuade* him to do so, the question arises where to draw the fine line between persuasion and manipulation? What are ethically acceptable ways of using technology to persuade people to change their behavior?

Philosophers have tried to answer these moral questions and tried to come up with initial ethical guidelines for the design and usage of PT. Amongst these attempts are approaches that look at PT from very different perspectives including 'post-phenomenological mediation theory', 'rhetoric', 'Rawlsian Liberalism', 'democratic technology design' and 'advertisement ethics' (to name just a few).[1] Despite these efforts, it is fair to state that the ethical reflection on persuasive technology has not yet succeeded to deliver a convincing framework for judging these difficult issues.

In the following essay I try to develop a coherent set of ethical guidelines that stems from the application of discourse ethics to persuasive technology. The

[1] See e.g. Johannesen (1992), Berdichevsky and Neuenschwander (1999), Baker and Martinson (2001), Fogg (2003), Verbeek (2006), Christensen and Hasle (2007), Davis (2010). The insightful article by Baker and Martinson comes closest to my own approach. However, Baker and Martinson only discuss 'persuasion' in the context of advertisement and do not consider persuasive technologies. Furthermore they do not link their ideas to an elaborated ethical theory, such as discourse ethics. While Davis is using a framework that might very well be linked to discourse ethics, she is mainly concerned with the *procedural* aspects of democratic technology design and does not discuss *material* norms that might be derived from a Habermasian point of view. I have sketched initial ideas to link discourse ethics with PT in Spahn (2010).

interesting aspect of PT is that they can be interpreted as a ‘communicative’ type of technology. They try to ‘persuade’ the user and thus establish a communicative relation between the designer and the user of the technology. In order to derive at these guidelines, I will interpret discourse ethics as a modern version of ethical rationalism and apply its insight to PT. I will distinguish between two elements of discourse ethics, both of which will lead to concrete suggestions for ethical requirements—stemming from the idea of the inherent normativity of speech-acts (2.1) and from the distinction between two types of rationality: communicative and strategic rationality (2.2). I will finally discuss objections to the idea of using discourse ethics in the context of PT and defend the claim that a discourse ethical framework is a very promising ethical framework given the communicative nature of PT (3).

Discourse Ethics and Persuasive Technologies

I would like to develop several criteria for the design and usage of PT that can be derived from the basic ideas of discourse ethics. Discourse ethics itself is probably one of the most interesting versions of a contemporary rationalistic, deontological approach in ethics. Choosing a rationalist framework to investigate PT poses a special challenge, since the rationalist tradition has always highlighted the importance of moral autonomy and well-considered ethical reasoning in making moral decisions about how to act. I will argue, that embracing a rationalist framework does not entail a rejection of PT. Rather this type of ethics offers interesting insights, about how to make these technologies meet our moral standards. The focus of the following paragraphs will thus lie in developing the main idea of this approach and in giving a first programmatic overview, that I intend to develop further in future work. In doing so, I will interpret discourse ethics as rationalistic ethics, which consists of a specific combination of two strands of ideas: a normative interpretation of analytic speech-act theory and a revival and modernization of continental German Enlightenment Rationalism (or Neo-Kantianism). I will therefore distinguish between four criteria that can be deduced from the ‘speech act’ elements of discourse ethics “The Elements of Speech-Act and PT” and three guidelines that stem from the wider social-neo-Enlightenment context of this theory “The Neo-Enlightenment Aspects of Discourse Ethics and PT”.

The Elements of Speech-Act and PT

One of the main inspirations in the development of early discourse ethics has been the speech-act theory by Austin and Searle.[2] Speech Act theory widens the philosophical attempts to understand the ‘meaning of words’ beyond the purely propositional elements of language. Ever since the linguistic turn early analytic philosophy dedicated much effort to detailed logical and epistemological analysis of

[2] Austin (1962), Searl (1969). For more recent analyses of speech-acts see Eemeren (1982), Asher and Lascarides (2003), Barker (2004), and Vanderveken (2009).

propositions and their truth conditions. Speech-Act theory overcomes this narrow focus, by including a reflection on the pragmatic dimension of language: it does not only take the 'propositional content' of speech into account, but puts an emphasis on the 'illocutionary act', which is a central element of every speech-act.[3] Austin thus famously analyzes 'how to do things with words'[4] and suggests a framework for understanding how speech acts can establish social facts like marriages, promises, orders and the like.

Discourse ethics emphasizes the normativity of the conditions for successful speech-acts. Whenever I utter a speech act, I have subscribed as it were to the rules of the 'game of communication' and other people expect me to follow these rules. If I do not use speech-acts correctly, my interlocutors might be irritated or even angry with me. Applied to persuasive technologies, one can thus assume that it must also adhere to some of these fundamental rules, if it is to be perceived as an attempt of 'persuasion' by the user. Users might be irritated if PT fails to meet these basic expectations. Furthermore, discourse ethics starts from the notion of an idealized concept of communication: it focuses on a rational discourse, whose aim it is to jointly search for and agree upon the truth of claims made in this discourse. Discourse ethics has thus two normative elements: it claims that every real discourse presupposes certain (normative) rules as necessary preconditions for the search of the truth, and it claims secondly that the outcome of a rational discourse has a normative validity: it is a rational consensus that is normatively binding as the result of a rational intersubjective deliberation process.[5]

In the context of this paper I cannot analyze the various claims of discourse ethics in detail, and I will not be able to discuss whether this normative interpretation can be a solid starting ground for establishing a contemporary neo-Kantian rationalist framework in ethics. Rather I would like to take the core ideas of discourse ethics as a starting ground and investigate, whether they can be applied to persuasive technologies.[6]

According to Jürgen Habermas, there are four implicit validity claims inherent in every intersubjective speech-act: a claim of comprehensibility (of the utterance), a claim of truth (of the propositional content), a claim of rightness or appropriateness (of the illocutionary element of the speech act) and a claim of truthfulness with regard to the intentions of the speaker (Habermas 1973). Whenever one of these claims is not fulfilled, a conversation may be doomed to failure or give rise to misinterpretations and misunderstanding (McCarthy 1996). If, however, these validity claims are preconditions of every successful speech-act, they must also be valid for 'persuasion' since persuasion is a speech-act. If the usage of persuasive technology can be interpreted as the attempt to persuade someone with the help of

[3] Searl (1975).

[4] Austin (1962).

[5] Sometimes discourse ethics is (wrongly) interpreted only as a procedural ethics. I agree, however, with Werner, who points out that discourse ethics is neither a domain specific ethics, nor a purely procedural theory (Werner 2002).

[6] For recent discussion of discourse ethics see Habermas (1993), Werner (2002), Kuhlmann (2007), Niemi (2008).

technology, these four validity claims may shed light on ethical design principles for this type of technology.

It seems that we can apply four prima facie guidelines that are relevant for the usage and development of persuasive technologies. In what follows, I will sketch those rules. Future case studies might show how fruitful this approach is for understanding and designing ethically acceptable persuasive technologies.

Comprehensibility

Every speech-act has to be 'comprehensible' in order to count as a potential element of communication. Incomprehensible utterances might at best be borderline cases, which cannot be seen as successful intentional attempts to transfer meaning. One can thus argue that comprehensibility is the most basic precondition for communication. A communication can only be successful if the utterances of the interlocutors are comprehensible.

If one applies this idea to PT it implies that the user must be capable of understanding the feedback given by the persuasive technology in question. Designers of PT must thus consider the question: Is this type of persuasive technology easy to understand?

Since PT mostly does not use 'speech' as a means to give feedback, the issue of comprehensibility becomes even more important. PT uses feedback to transfer a message: a red light might signal, that 'something is not right', while a 'smiling face' in the display of a RSI Guard can be read as 'you are doing fine'. The type of feedback used varies wildly in PT, from 'very simple' to quite advanced. Very uncommon and advanced is e.g. the feedback used in the 'Persuasive Picture Frame' (Obermair et al. 2008). This picture frame is a device that monitors poor sitting habits and bad sitting posture of computer users. It displays a changing portrait of a person that the computer user loves or likes. This portrait varies in facial expressions from 'angry' to 'happy', depending on the sitting posture of the user. The aim is to motivate the user to adapt a healthy sitting posture. Since this feedback mechanism relies on human facial expressions, it is very easy to interpret by the user. Most feedback mechanism used in PT are, however, much less advanced and offer thus many options for misunderstandings. With regard to comprehensibility it might therefore be helpful to work with symbols, colors or signals, that do either have a strong conventional or a strong visually fixed meaning. In short: the used feedback mechanism should not lead to easily foreseeable misunderstandings.

An example of misinterpretation might be the Household Energy Saving Experiment that used PT in a social context (Bang and Jonsson 2007). The idea of this experiment was to give social feedback about energy consumption in a community of neighbors via different feedback mechanisms. One way of giving this feedback was placing a lamp in one window of the participating households. This lamp signaled visible to the neighbors how each household was doing with regard to their energy saving efforts. It did so by emitting either a green or a red light. 'Red' was immediately understood as signalizing a 'bad' consumption pattern. Thus using this feedback mechanism did at first glance fulfill the requirement of comprehensibility. After the experiment, however, worries were expressed by the participants, that this red-light was only giving feedback about the energy consumption relative to the previous pattern of the same household (e.g. ‘this household is doing worse than last week’), thus communicating a ‘wrong’ signal, if it was interpreted as an ‘absolute’ statement (e.g. ‘this household is doing bad on a fixed scale’). PT thus has to use clear and straightforward ways of communicating feedback, especially if it is used in a social context to avoid misunderstandings and fulfill the requirement of comprehensibility.

Another example of a PT that might suffer from a lack of ‘comprehensibility’ is the Eco Ambient Awareness Dashboard, described by Kim et al. (2010).[7] They designed two Mac OSX eco-visualization widgets, which were intended to raise awareness about energy consumption. One was based on accurate numerical feedback; one was based on a nice graphical visualization. The researchers concluded that the second widget was more effective in ‘persuading’ the user, since it was easier to connect with its message in an intuitive way, which required not much cognitive effort. This widget displayed a graphic animation of a coral reef. The coral reef was ‘healthy’ and full of plants and fish, incase the user did not have his computer turned on for long in the ‘idle mode’. If the user would, however, have a lot of idle computer time, the coral reef would slowly change to represent an unhealthy and dying coral reef. This eco visualization was ‘comprehensible’, if it was understood as representing the efficiency of one’s computer-usage. But it was not comprehensible if read as reflecting total energy usage. Somebody who uses his computer very often, but with little to no idle time, will experience the joy of having a ‘healthy coral reef’—despite using much energy in total. However, somebody who uses his computer very little but with a worse ‘idle-time’/‘usage-time’ ratio, might have a ‘dying coral reef’, even though he might be using less energy in total. This shows that designers have to carefully consider, whether the users will not only ‘comprehend’ the evaluative element of the metaphoric ‘visualization’, but also whether they will be able to understand, *what exactly* is visualized.[8]

Comprehensibility becomes even more relevant as an ‘ethical’ aspect of communication, when it is the explicit secret intention of the designer (or the person who implements a PT) to give incomprehensible information, with the hope that he will be misunderstood. This is a common and often criticized tool in dodgy yellow press or in misleading advertisement spots: choosing a statement, which is easily misunderstood, without claiming something that is strictly speaking false.[9] This leads us to the second validity claim of speech-acts.

[7] Kim et al. (2010).

[8] This implies that sometimes designers might have to face a difficult trade-off. A PT might be easier to *use*, if it gives very simple graphical feedback (ie. feedback that can be processed very quickly by the user without too much cognitive attention). However this ease of use might in some cases lead to the danger that it is also easier to *misunderstand* the feedback.

[9] A famous example is the joke about two persons A and B who compose a blackmail letter. A is dictating the letter and B is writing it down. When faced with the accusation that they were responsible for this letter, A says: ‘I would never write such a letter’, while B claims: ‘How dare you think, I could formulate such an ill-spirited letter.’ Strictly speaking they are both not lying.

Truth

From a philosophical perspective the truth claim can be regarded as the most important validity claim of speech-acts. Persuasive technologies communicate with the user via various feedback mechanisms. The truth requirement demands thus that this feedback mechanism is actually giving true information. An energy saving device, like e.g. the Wattson Energy Meter, that gives a feedback about the expected annual costs of a given energy consumption pattern, should of course display the real price—even though the 'persuasive effect' might be higher, if the device 'lies' and exaggerates the costs a bit.

What is at stake here, is the key question of how to evaluate rhetorical elements of speech-acts from an ethical perspective: It is ethical to convince someone, no matter whether what one says is true, or is real rhetoric—as Plato would claim[10]—linked to finding that aspect of *truth* that is able to convince somebody? One can argue, that from an ethical standpoint PT should highlight from the wide range of possible factual information that part, that can best be used to convince somebody to change his behavior, and thus it should always be linked to a verifiable propositional content.

Baker and Martinson's (2001, p. 154) illustrate this point as follows: "One who communicates false, misleading, or deceptive information in a serious circumstance, from a moral philosophy perspective, does wrong, because to do so perverts the very purpose of speech (communication)." In human communication we rely upon the idea that the other one is not trying to deceive us. Also in Human-Technology interaction, this idea becomes relevant and leads to the aspect of 'truthfulness'.

Truthfulness

In human communication, the property of a speaker to tell the truth in a given context is called the 'honesty' or 'truthfulness' of the speaker. Honesty is a central element in establishing trust in human relations. Can there also be honesty or truthfulness in PT? One of the questions that might arise is the concern that it is not clear who the 'author' of the persuasion is. Is it the persuasive technology itself, is it the designer of the PT, is it the company or institution that implements a given PT? If e.g. the usage of smart meters is prescribed by national law to decrease energy consumption, it might be plausible to argue that the subject of the persuasion is the government. If, however, a user freely decides to install a RSI guard on his computer it seems that he himself it the initiator of the persuasion. In "Discourse Ethics and the Communicative Nature of PT: Concluding Remarks" I will discuss the question in more detail, in how far technologies can be 'persuasive' at all, given that they do not have 'intentions' or 'mental states'. Thus, properly speaking they can also not be 'truthful' in the same sense that applies to humans. For now it might be enough to interpret 'truthfulness' in the case of PT as reliability and accuracy of feedback mechanisms. Even though these questions are certainly relevant, and will

[10] See the *Gorgias* and *Phaidros*. For an analysis of Greek rhetoric and its relevance for modern strategies of persuasion compare Marsh (2001).

be dealt with later, one can thus try to locate the ‘truthfulness’ in the feedback mechanism of the PT itself. The best way to meet this validity claim might then be to ensure that it only displays true and accurate information. As in human communication, ‘honesty’ (a property of a person) and ‘telling the truth’ (a property of a proposition) are also in PT closely related.

Fogg (2003) has argued with regard to this point in a similar vein, that computers are powerful persuaders precisely because they represent in the public opinion a kind of mathematical precision that may serve as a role model of ‘true and accurate’ feedback. It has to be noted that, however, this might only be true with regard to the general image of computers as powerful persuaders. As persuasive technology becomes ‘ambient’ and moves away from the personal computer into the environment of the user, this assumption might no longer be true.

From an ethical standpoint one might even construct a Kantian argument for designing and creating PT according to principles of honesty and trustworthiness. If a given PT turns out to be not trustworthy by ‘twisting the truth’ or giving false or exaggerated feedback, it may be a powerful persuader in the short run (as long as this ‘dishonesty’ stays unnoticed). If the user, however, learns that this PT has as it were ‘cheated’ on him, he might be less willing to trust the specific device; he might even develop a general mistrust in PT as a whole. In the long run PT has to be truthful or else it won’t be trusted. To use the analogy of lying: lying might sometimes be more effective to achieve a certain goal and it might thus give a strategic advantage for the person who lied. But lying also undermines the trustworthiness of the person who lies. It therefore—as discourse ethics rightly claims—undermines a general presupposition of human communication, and can for that reason not be a stable strategy that works *in the long run*. An example could be the old habit of putting breakfast cereals into huge packages, suggesting that the consumer would by a big quantity of them. Companies abandoned this strategy as it became visible that consumers understood this ‘cheat’ and where disappointed by this attempt of untruthful persuasion.

Appropriateness

The criteria of appropriateness concerns the question, whether a given speaker is entitled to the illocutionary element of the given speech act. A priest may utter the sentence: ‘I declare you husband and wife’. By doing so he establishes at the same time a new social fact. But this is true only if uttered under very specific circumstances by a priest, who is entitled to do so. Many speech acts like ‘orders’, ‘promises’, ‘excuses’ and the like can be judged ethically according to their ‘appropriateness’.

With regard to PT the question of appropriateness arises on two levels. One may ask whether PT is an ‘appropriate’ means to reach the end in question. If I want to change the energy consumption behavior of car drivers, there are many options to reach this end. Passing a speed-limit, adding taxes on fuel, convincing people to care more for the environment, etc. are all different options to try to reach this aim. Installing ‘persuasive technology’ in cars (e.g. by eco visualization feedback) just adds one more option to this range. It is clear that speech-act theory as such cannot answer this very global question. In “The Neo-Enlightenment Aspects of Discourse Ethics and PT” I will try to shed more light on this question, by turning to the second ingredient of discourse ethics: its neo-enlightenment background.

A second way to consider the criteria of ‘appropriateness’ concerns the more fundamental question, whether a given technology should be labeled ‘persuasive technology’ at all. Many technologies, that try to bring about behavior change, can be seen as attempts of ‘manipulation’ or even ‘coercion’ rather than ‘persuasion’. There is a danger, that the label of ‘persuasive technology’ might be just a nicer way to refer to instruments of propaganda or manipulation. In order to address this question, it is once again necessary to go beyond the elements of speech-act theory. In looking at the second element of discourse ethics, I will suggest how one might address these complex issues within this ethical framework.

There is also a third way to interpret the question of appropriateness. One might ask the question, whether *any technology* should be labeled ‘persuasive technology’, since persuasion refers to an act of human communication.[11] Can technologies, that do not have intentions or mental states thus be persuasive at all? I will address this third question in “Discourse Ethics and the Communicative Nature of PT: Concluding Remarks”, when I discuss general objections to the suggestion to use discourse ethics for the field of PT.

Even though these questions will certainly need closer elaboration, it can be seen that the speech-act element of discourse ethics proofs to be a very fruitful starting ground for structuring the debate on ethical issues of PT and for developing an ethical framework for them. After all ‘persuasion’—with or without the usage of technology—remains a communicative act, an act in which e.g. the designer of a PT wants the user to change his behavior to reach a desired aim. If there are general ‘rules of the game’ of speech-acts that are morally relevant in general, they might as well be applied to the concrete design and usage of PT.

The Neo-Enlightenment Aspects of Discourse Ethics and PT

It should have become clear that the application of the above mentioned four criteria to PT leads to an interesting way of looking at this type of technologies. But one has to admit, that these criteria are still on a very abstract level (as they can be applied to all types of communication) and may thus not yet lead to more substantial rules for what is and what is not allowed in the specific contexts of *persuasion*. Or to put it in another way: the general ethical rules for the ‘game of communicative interaction’ might differ from the specific rules of the ‘game of persuasion’. However, one can argue, that the game of ‘persuasion’ is a subset of the language game of ‘communication’, and the last paragraph illustrated, in how far the general validity claims of communication can be made fruitful for persuasive technologies.[12]

[11] I thank one of the anonymous reviewers for directing my attention to this third question.

[12] Jilles Smids pointed out to me, that one might even turn this argument around and claim, that if a persuasive technology does not adhere to those minimal conditions, it should not be labeled a ‘persuasive technology‘, but rather a ‘behavior-influencing’ technology. Discourse-ethics might thus provide normative criteria for a definition of persuasive technology.

If one, however, also takes the more general background assumptions of discourse ethics into consideration, one may derive at further guidelines for the design and usage of PT that might indeed be more substantial (and thus also more controversial), than only focusing on the general aspects of all communicative relations. In the remaining part, I thus develop some general principles that can be deduced from this second element of discourse ethics.

Considering the broader context of discourse ethics in the writings of Jürgen Habermas and Karl-Otto Apel, one is faced with the attempt to establish a critical theory of the social reality. This general theory of social reality forms the background in which discourse ethics is embedded. An important idea is in this context the well-known Habermasian distinction between different types of rationalities, which he links to different types of knowledge and to different types of social institutions. The main idea that is relevant for our context is the distinction between communicative rationality and strategic rationality.[13]

Strategic rationality is the type of rationality that seeks the adequate means for a given end. It thus mirrors Kant's definition of hypothetical imperatives and is concerned with *knowing how* to achieve one's goals. In the context of critical theory this type of knowledge is seen to be rooted in man's need to dominate nature, which is necessary for his survival. Strategic rationality is regarded to lie at the heart of technological knowledge and aims at an exercise of power: power over nature or power over other human beings.[14] Opposed to this type of knowledge is the communicative rationality. It is linked to the social nature of man as a 'political animal' (Aristotle), and aims at insight in ethical values and in emancipation from unjust power relations. It mirrors Kant's interest in categorical imperatives, which form for him the core of ethical knowledge. While strategic knowledge implies an asymmetrical relation, communicative rationality starts from the general ideal of symmetry: every voice is to be taken serious in the discourse about morality and only the force of the better argument counts—independent of external factors about the person who utters the argument.

Critical theory tends to see, roughly speaking, strategic rationality as an (sometimes necessary) evil, and communicative rationality as the ethical ideal. It seeks thus to eliminate asymmetrical exercises of strategic rationality as often as possible and complains about the fact that our life-world, which should be structured according to the rules of communicative rationality, gets more and more 'invaded' by the logic of strategic rationality.[15] If one, however, wants to add

[13] Habermas (1984–1987). Apel (1984).

[14] Habermas (1971). This is the point where Habermas comes closest to the technik-critique of Heidegger and Adorno.

[15] The relation between strategic rationality and communicative rationality is more complex in Habermas than this simplified sketch suggests. However, I cannot discuss the details here. There are of course many cases, in which the exercise of strategic rationality is morally justified and unproblematic. But in order to justify or determine, whether or not a given case of strategic rationality is morally unproblematic, one needs consensual norms and thus communicative rationality (cfr. the discussion of this relation in McCarthy (1996, p. 17ff., esp. p. 29). In the context of justification there is thus a primate of "communicative rationality" over "strategic rationality".

'persuasive rationality' to the picture one is faced with the problem that persuasion seems to fall exactly in the middle between the two categories.

Type of rationality	Aim and method	Type of relation
Communicative rationality: 'convincing'	Shared values, behaviour change through insight	Symmetry
Persuasive rationality: 'persuasion'	Social values, behavior change through 'nudges'	Between symmetry and asymmetry
Strategic rationality: 'manipulation'	Know-how, exercise of power	Asymmetry

It seems that from the perspective of discourse ethics, all other attempts to arrive at a change of behavior seem ethically inferior to the exercise of communicative rationality. However, since 'persuasive rationality' seems to be neither purely communicative nor purely strategic, it is not yet clear how to deal with it. I would like to suggest that the best way to look at PT from a discourse ethics perspective is to acknowledge the intermediate position of persuasion and try to make it as less asymmetrical and strategic as possible. Therefore, I will suggest three ways of doing so, that might be suitable to serve as *prima facie* design-guidelines for PT.

Looking at the general set-up of critical theory it seems that there are mainly two ethical concerns, that are related to strategic rationality, that could thus still be applied (and be it only too a lesser degree) to the idea of a 'persuasive rationality'. The one concern has to do with the asymmetric relation, that is inherent in strategic rationality and still plays a role in PT. The second concern has to do with the fact that strategic rationality is often a form of manipulation, which does not take the autonomy of the other serious. The following guidelines are meant to take care of these worries.

These guidelines are:

G1: Persuasion should be based on prior (real or counterfactual) consent.

Variations of this principle have been discussed in literature. In an early article on the ethics of PT, Berdichevsky and Neuenschwander (1999) present what they call a 'Golden Rule' for the design of PT. It reads:

> The creators of a persuasive technology should never seek to persuade a person of something they themselves *would not consent to* be persuaded to do.

The idea behind this rule is to eliminate the asymmetry of the persuasive situation. One may only use persuasion, if one is willing to be subject to the very same type of persuasion for the very same end. As presented in their paper, however, this rule has some serious shortcomings.

One objection, that the authors discuss themselves, is the example of a person, who wants to persuade someone else, even though he does not want to be persuaded himself. Think about a father who does not want his son to smoke, even though he himself is not willing to be persuaded to quit smoking. In a similar line a far more serious objection could be made in the case of someone, who is willing to accept an immoral form of persuasion for an immoral aim, and might thus be willing (and by the rule of Berdichevsky and Neuenschwander even morally entitled) to use it on others. Think about a racist who intends to persuade others (and himself) to accept racist ideas.

The reason why Berdichevsky and Neuenschwander cannot plausibly reject these objections lies in the fact that they take *factual consensus* as starting ground. Discourse ethics, however, is aiming at *ideal rational consensus* and is thus not seeking for what people actually agree upon, as they might indeed agree on very immoral ideas. Rather the focus is on what a rational subject *should* agree upon, if he is only taking the better arguments (and not his irrational desire to smoke or to value other races lower than his own race) as decisive reason for action. An analysis of this 'Golden Rule' of persuasion from a counter-factual ethical theory—as discourse ethics is—is thus better able to counter these objections than a factual interpretation of consensus.

The distinction between 'ideal counterfactual consensus' and 'real factual informed consent' can thus be very helpful for the design and usage of PT, if one applies it in the following way. For a *philosophical debate* about the moral evaluation and justification of PT the idea of a counter-factual ideal consensus should be the proving ground for justifying which attempt of persuasion is ethically acceptable. For a concrete *technological implementation* of PT in everyday life situation, however, a real factual informed consent might be enough to strive for. In both cases the general idea is to limit the asymmetry of the 'persuasive situation' by linking it to a prior symmetrical relation. If I agree to be subject to a persuasion, this persuasion becomes closer to an attempt of convincing, than if I do not agree to it—in the later case the same type of persuasion might be experienced as an attempt of 'manipulation' rather than 'persuasion'.

G2: Ideally the aim of persuasion should be to end the persuasion.

Also this rule is related to the idea of limiting the asymmetry inherent to PT and can be placed in the broader context of critical theory. This guideline starts from the obvious difference between 'education' and 'manipulation'. Both types of action presuppose an asymmetrical relation, but from an ethical standpoint both are to be judged very differently. The aim of education is namely to create an autonomous person, who in the end will be able to educate himself and will no longer need guidance. The aim of manipulation is on the contrary to keep the other in a state of dependence, or even worse, to make him more dependent than he was in the first place. Education aims at overcoming the asymmetry, manipulation aims at fostering it. If one applies this idea to PT, one can argue that persuasive technology should follow the principle of 'education', rather than the structure of 'manipulation', i.e. it should make itself superfluous.

One example could be a gas-pedal that gives a mechanic feedback of resistance if the driver pushes it too hard and thus is about to waste energy. Initially this behavior of the paddle will irritate the driver, but after a while he will 'learn' the right way to accelerate and he will feel the ‘counterforce’ on fewer and fewer occasions. In a similar line a computer program that forces the user to make work breaks after typing for a long period of time, might be less experienced as paternalistic, if the user learns to think about the breaks himself, and thus does no longer experience the program as an external force and might at one point even delete it, if the new behavior has become his ‘second nature’. Of course this is a very general guideline that might not be applicable to all types of PT, but in designing persuasive technologies it will be helpful to have the analogy of ‘education’ vs. ‘manipulation’ in mind and to think about the implementation of a learning effect, and thus about a potential ‘ending’ of the asymmetric nature of persuasion.

G3: Persuasion should grant as much autonomy as possible to the user

This final guideline tries to make the ‘persuasion’ more compatible with the ideal of ‘communicative rationality’, which focuses on autonomous decisions. This guideline is very broad and needs a more specific elaboration. But if one starts from the discourse ethical ideal of ‘convincing’ one might be able to get a better understanding about what it means to grant more ‘autonomy’ to the user: In a communicative situation both interlocutors are on a level of symmetry and can exchange arguments. Each one has thus the ability to have an influence on the other. PT is, however, ‘one-sided’ with regard to this aspect of communication. It is often purely designed “in one direction”: it aims at changing the attitude or behavior of the user. A way to limit this asymmetry could be to let the user choose amongst various options, i.e. to inbuilt plurality and choice in the PT. To use again the example of a RSI Guard, one can imagine a program that lets the user decide how ‘harsh’ the persuasion will be. Will the computer simply refuse to take any inputs during a ‘break’? How easy will it be to override the persuasion? How easy will it be to change from strict coercion to mild suggestion?, etc. In giving the user the ability to make autonomous decisions about these questions beforehand, one may further limit the asymmetry and the paternalism that seems to be so deeply linked to PT.

With regard to autonomy, it is furthermore important to distinguish whether the user persuades *himself* with the PT, or whether the PT is used to persuade him *by someone else*. Cases of self-persuasion are most likely cases, in which the user already shares the value in question and uses the PT only to overcome a ‘weakness of the will’. These cases might be morally less problematic than using PT to persuade others. This might include cases in which the technology targets not only a behavior, but an attitude change. Future research on the relation of autonomy and PT will have to take this distinction into consideration.

Discourse Ethics and the Communicative Nature of PT: Concluding Remarks

I tried to demonstrate, how guidelines for the design and usage of persuasive technologies could be derived from the framework of discourse ethics. In the remaining section I would like to discuss some objections to this approach and conclude by summarizing the main ideas of the paper.

(A) The first objection to the application of discourse ethics to PT is that persuasive technologies do not fall under 'communication' between humans, and should thus not be subject to the same rules as human communication. There is a difference between inter-human communication and the usage of PT. This objection is a very basic one. It questions the very idea of 'persuasive technologies'. The argument runs as follows: Only humans can really persuade each other and communicate with each other. Technology is not capable of speech, mental states, intentions and other crucial elements needed for an act of 'persuasion.' If technology cannot be persuasive properly speaking, an ethical framework designed for human communication cannot be applied to technology.

The argument seems plausible at first glance. However, it makes certain assumptions that can be questioned. The two problematic assumptions are (1) 'technology cannot persuade' and (2) 'discourse ethics can only be applied to communicative relations between humans (and not to the relation between humans and technologies)'. If both assumptions were true, discourse ethics could not generate an ethical framework for PT. I would like to argue that both assumptions are wrong.

There are two ways to counter the assumption that 'technology cannot persuade in a proper sense', and that 'therefore discourse ethics cannot be applied'. The first way is to accept that technology cannot persuade properly speaking, but to defend the idea, that this does not imply that discourse ethics cannot be applied, since there is still a persuasion between *human* agents going on (1). The second option is to argue that technology is indeed capable of 'persuading humans', even if of course in a different and more limited sense, than real humans can acheive (2).

(1) If one insists that persuasion presupposes human agency (and thus mental states, intentions and the like), then Pt can at best be a *mediated* form of persuasion. In this interpretation of PT the actual persuader is the designer of the PT or the person (or group of persons) who implemented the PT. If e.g. the Dutch government prescribes 'Smart Meters' to persuade people to use less energy, the author of the persuasion is the 'Dutch Government' and not the technology itself. This interpretation regards the usage of the words 'persuasive technology' only as a metaphorical way of speaking. But even if one decides to embrace this interpretation, one can still apply the two elements of discourse ethics that have been discussed in the previous section. The reason is that there is still an intersubjective 'persuasion' going on that is *only mediated by the usage of technology*. Since PT aims thus at 'persuading' the user to change his attitude, belief and/or behavior (Fogg 2003), they can in this interpretation be regarded as a mediated form of communication between a 'persuader' and a recipient of the persuasion. Since discourse ethics deals with general rules of communication it is thus worthwhile to look at PT from a discourse ethical perspective. Persuasive communication forms a subpart of general communication, and persuasive communication via technology again forms a subpart of persuasive communication.

Of course one can reasonably expect more specific rules for each 'sub'-domain. Since persuasion, mediated by technology, does differ from 'not mediated' attempts of persuasion, one will have to develop *special* rules that account for this difference as well. This is an interesting field for future research. In this context, however, I intended to focus on the *general* rules of the relation of ‘persuasion’ that can be applied to all types of persuasion (mediated or not). These general rules of communication should be valid also in the sub-types. To illustrate this I argued that particular validity claims (such as comprehensibility, truth, truthfulness and appropriateness) should be part of any *ethical* attempt of persuasion.

(2) One might try to go even further and defend a broader understanding of ‘persuasion’ and argue that the ‘persuasive relation’ should not be limited to purely intersubjective human relations. Some philosophers of technology have argued that the human-things dichotomy is too sharp and should be overcome. According to this position it is wrong to see only human agents as ‘active prime movers’ and technological tools only as passive and neutral tools. Rather technologies posses ‘agency’ or even ‘intentionality’ (e.g. Ihde 1990; Verbeek 2005, 2008) and should be interpreted accordingly. Even though I am quite critical of these approaches myself (Peterson and Spahn 2011), one can see, how embracing this interpretation of technological artifacts allows the extension of a framework like discourse ethics to human-technology relations as well. If the distinction between human ‘agents’ and technological ‘tools’ gets blurred, one will also have to consider broadening the domain to which ethical theories such as discourse ethics can be applied.

In this paper I cannot investigate the benefits and limits of this interpretation of technological artifacts. However, I want to point at a third way to defend the application of discourse ethics, that relates to human psychology. Humans tend to ascribe ‘agency’ to even very abstract non-living things.[16] They do so even more in the case of technologies that use social feedback or social ‘clues’ (Mayer et al. 2003)—like PT often do. From a philosophical perspective, one might of course still criticize this human tendency. One can of course argue that humans *should* not attribute agency to PT, even if they are inclined to do so. But the fact that they tend to do so gives an additional reason to apply discourse ethics: If humans tend to engage with ‘technological persuaders’ in a similar way and with similar expectation as in the case of ‘real’ agents, designers should take this into account. This implies reflecting on (discourse) ethical aspects of communication as well.[17]

(B) A second objection to the general approach of this paper might be to insist that the discourse ethical framework was developed to create guidelines for a specific type of communication. Discourse ethics aims at giving guidelines for philosophical and political “discourses”, why then should it be applied to PT? The aim of PT is not to find out the truth in philosophical matters or to establish a justified political decision, but to change the behavior of a user in a very concrete context. The idea behind this objection is more specific than the first objection. While it grants the fact, that PT can be regarded as establishing a communicative

[16] The classical study in this field is by Heider and Simmel (1944). A more recent discussion of this ‘anthromorphism’ can be found in Wayetz et al. (2010).

[17] Nickel (2011a, b) has recently argued from a philosophical perspective that human attitudes like ‘trust’ can be applied to technological systems in a meaningful way. Like ‘persuasion’, trust originally refers to a human–human relation. The type of arguments that Nickel uses to defend the application of ‘trust’ and ‘trustworthiness’ to artifacts are similar to the arguments presented in this paper.

interaction, it nevertheless insists that PT establish a certain type of communicative interaction: a type for which the discourse ethical framework might not be valid.

This objection, however, presupposes a purely formal (or procedural) interpretation of discourse ethics. This interpretation might be widespread within the domain of applied ethics, such as ethics of technology and democratic technology design, but it does in my view not capture the heart of discourse ethics. In this procedural interpretation, discourse ethics develops rules for establishing a framework for an intersubjective discourse, but it does not answer any material questions. In this interpretation, there can be no substantial answers to ethical questions from a discourse-ethical perspective, except the general advise to 'talk' about the issue in question, while adhering to certain discourse rules, as e.g. the absence of power-relation and the willingness to let only the better argument count. As said, I am not convinced that this procedural interpretation covers the complete intentions of discourse ethics.[18] But even if it were correct, one could still defend the application of discourse ethics to PT. One would simply have to argue, that discourse ethics only establishes material rules for procedures of 'communication' (and thus not predetermines the outcomes of this communications). But if PT can be interpreted as communicative technologies, then the norms of communication should also be applied to this type of communication. In short: if the objection is that discourse ethics can only establish procedural norms (but no material ones), then the answer is, that procedural rules are all that is needed to evaluate PT and the communicative aspects of PT. In this paper, I tried to go beyond a procedural interpretation of discourse ethics by highlighting also the neo-enlightenment aspect of this approach.

(C) This leads to a final remark that moves beyond the scope of this paper: The question remains open, *under which circumstances* it is ethically acceptable to choose 'persuasion' rather than other options of behavior change like 'convincing' or 'coercion'. This paper is only concerned with the development of a moral version of 'persuasive technologies'. If the aim is to reach a behavior change, then 'persuasion' is of course only one strategy amongst others. In some situations the only moral strategy might be "convincing through rational arguments", in other situations even "coercion" or "manipulation" might be ethically justified. We might e.g. coerce someone to pay taxes. Or we might want to manipulate an evil person, so that he gives up his evil plans, etc.

In this paper I did not address the question, when to chose which strategy. The argumentation of this paper is thus conditional: if one wishes to use PT, one should consider certain prima facie guidelines. I did not investigate, *whether* one should use PT in a certain context or for a specific task. Within the framework of discourse ethics, there is of course place to justify the use of coercion, persuasion or other strategies for certain situations. The task to establish criteria, when to use persuasion, rather than other means of influences, is however different from the scope of this paper and requires further investigation. The task here was to argue, that *if* one uses PT at all, certain ethical criteria and prima facie guidelines should be

[18] See Werner (2002) for a defense of a material interpretation of discourse ethics.

fulfilled. I leave the question *when* to use PT (i.e. in which situation and for which purposes) to future analysis.

Summary

This essay aims at giving an overview of how a discourse ethical framework for PT might look like. It does so by distinguishing between two elements of discourse ethics: the speech-act aspect and the social embedding of the theory. The two main ideas are the insight that 'persuasion' has to meet the implicit normative presupposition of every act of communication: criteria of comprehensibility, truth, truthfulness and appropriateness can thus be applied to PT. Furthermore PT can be interpreted as holding the middle between strategic and communicative rationality. In order to assure that this type of communication is ethically sound, one would have to try to limit the strategic elements and foster the implicit elements of 'communicative' actions. Three prima facie guidelines have been suggested to overcome or at least limit the often asymmetric nature of persuasion.

The ideas developed in this essay are a first attempt to make discourse ethics fruitful for the debate about PT. Further research is needed to come up with more precise guidelines and criteria and to develop this approach further into a full-fledged ethical framework for PT. However, it should have become plausible, that discourse ethics indeed is a very promising starting ground for the search of guidelines in the field of PT.

References

Apel, K.-O. (1984). Das Problem einer philosophischen Theorie der Rationalitätstypen. In H. Schnädelbach (Ed.), *Rationalitäẗ* (pp. 15–31). Frankfurt a.M.: Suhrkamp.

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge: Cambridge University Press.

Austin, J. L. (1962). *How to do things with words*. Oxford: Clarendon.

Baker, S., & Martinson, D. L. (2001). The TARES test: Five principles for ethical persuasion. *Journal of Mass Media Ethics, 16*(2), 148–175.

Bang, M., & Jonsson, L. (2007). Design of a social web service and ambient interface for energy conservation and social feedback in a neighboorhood. In: M. Tscheligi et al. (Eds.) *AMI09*. Salzburg: ICT&S Center.

Barker, S. (2004). *Renewing meaning: A speech-act theoretic approach*. Oxford: Oxford University Press.

Berdichevsky, D., & Neuenschwander, E. (1999). Toward an ethics of persuasive technology. *Communications of the ACM, 42*, 51–58.

Christensen, A.-K., & Hasle, P. (2007). Classical rhetoric and a limit of persuasion. In Y. de Kort, W. IJsselstein, C. Midden, B. Eggen, & B. J. Fogg (Eds.), *Persuasive technology. Second international conference on persuasive technology* (pp. 307–310). Palo Alto, CA: Revised Selected Papers.

Davis, J. (2010). Generating directions for persuasive technology design with the inspiration card workshop. In T. Ploug, P. Hasle, & H. Oinas-Kukkonen (Eds.), *Persuasive technology, PERSUASIVE 2010* (pp. 262–273). Berlin: Springer.

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do. The Morgan Kaufmann series in interactive technologies*. Amsterdam, Boston: Morgan Kaufmann Publishers.

Habermas, J. (1971). *Knowledge and human interests*. Boston: Beacon.

Habermas, J. (1973). Wahrheitstheorien. In H. Fahrenbach (Ed.) *Wirklichkeit und Reflexion* (pp. 211–265). Pfullingen: Neske.

Habermas, J. (1984–1987). *Theory of communicative action*, 2 Vol., Boston: Beacon Press. (translation of Habermas, J.: Theorie des kommunikativen Handelns, 2 Vol., Frankfrut a.M.: Suhrkamp 1981).

Habermas, J. (1993). *Justification and application. Remarks on discourse ethics*. Cambridge: MIT Press.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behaviour. *American Journal of Psychology, 13*, 1944.

Ihde, D. (1990). *Technology and the lifeworld*. Bloomington: Indiana University Press.

IJsselsteijn, W. (2006). *Persuasive technology: First international conference on persuasive technology for human well-being, PERSUASIVE 2006, Eindhoven, The Netherlands, May 18–19, 2006: Proceedings*. Berlin, New York: Springer.

Johannesen, R. L. (1992). Perspectives in ethics in persuasion. In C. U. Larson *Persuasion, reception and responsibility* (pp. 28–53). Boston, MA: Wadsworth, 2006.

Kim, T., Hong, H., & Magerko, B. (2010). Design for persuasion: Toward ambient eco-visualization for awareness. In T. Ploug, P. Hasle, & H. Oinas-Kukkonen (Eds.), *Persuasive technology 2010* (pp. 106–116). Berlin: Springer.

Kuhlmann, W. (2007). *Beiträge zur Diskursethik. Studien zur Transzendentalpragmatik*. Würzburg: Königshausen & Neumann.

Marsh, Ch. W. (2001). Public relation ethics: Contrasting models from the Rhetorics of Plato, Aristotles and Isocrates. *Journal of Mass Media Ethics, 16*(2), 78–98.

Mayer, R. E., Sobko, K., & Mauton, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology, 95*(2), 419–425.

McCarthy, Th. A. (1996). *The critical theory of Jürgen Habermas*. Cambridge, MA: MIT Press.

Nickel, P. J. (2011a). P.Trust in technological systems. In M. J. de Vries, S. O. Hansson, & A. W. M. Meijers (Eds.) *Norms and the artificial: moral and non-moral norms in technology*. Springer, forthcoming.

Nickel, P. J. (2011b). Ethics in e-trust and e-trustworthiness: The case of direct computer-patient interfaces. *Ethics of Information Technology*, forthcoming.

Niemi, J. I. (2008). The foundations of Jürgen Habermas' discourse ethics. *The Journal of Value Inquiry, 42*(2), 255–268.

Obermair, Ch., Reitberger, W., Meschtscherjakov, A., Lankes, M., & Tscheligli, M. (2008). perFrames: Persuasive picture frames for proper posture. In H. Oinas-Kukkonen, et al. (Eds.), *Persuasive technology, lecture notes in computer science* (pp. 128–139). Berlin: Springer.

Peterson, M. & Spahn, A. (2011). Can technological artifacts be moral agents, *Science and Engineering Ethics*, forthcoming.

Searl, J. (1969). *Speech acts*. Cambridge: Cambridge University Press.

Searl, J. (1975). A taxonomy of illocutionary acts. In K. Gunderson (Ed.), *Language, mind and knowledge* (pp. 344–369). Minneapolis, MN: University of Minnesota Press.

Spahn, A. (2010). Persuasive technology and the inherent normativity of communication. In P. Hasle, et al. (Eds.), *Persuasive 2010* (pp. 21–24). Oulu: University Press.

van Eemeren, F. H. (1982). *Speech acts in argumentative discussions*. Cinnaminson, NJ: Foris Publications.

Vanderveken, D. (2009). *Meaning and speech acts, vols. I and II*. Cambridge: Cambridge University Press. (^{1}1991).

Verbeek, P.-P. (2005). *What things do: Philosophical reflections on technology, agency and design*. Pennsylvania: Pennsylvania State University Press.

Verbeek, P. P. (2006). Materializing morality: Design ethics and technological mediation. *Science, Technology and Human Values—special issue on Ethics and Engineering Design*, 31/3, 361–380.

Verbeek, P.-P. (2008). Obstetric ultrasound and the technological mediation of morality: A postphenomenological analysis. *Human Studies, 31*, 11–26.

Wayetz, A., Epley, N., & Cacciopo, J. T. (2010). Social cognition unbound: Insights into anthromorphism and dehumanization. *Current Directions in Psychological Science, 19*(1), 58–62.

Werner, M. H. (2002). Diskursethik. In M. Düwell, et al. (Eds.), *Handbuch ethik* (pp. 140–151). Stuttgart/Weimar: Metzler.